GRASP Laboratory Technical Report #414, University of Pennsylvania. 30 May 1997.
Available at: ftp://www.cis.upenn.edu/pub/eero/buccigrossi97.ps.gz

Image Compression via Joint Statistical Characterization in the Wavelet Domain

Robert W. Buccigrossi
GRASP Laboratory
Dept. of Computer & Information Science
University of Pennsylvania
Philadelphia, PA 19104
butch@grip.cis.upenn.edu

Eero P. Simoncelli
Center for Neural Science, and
Courant Inst. of Mathematical Sciences
New York University
New York, NY 10003
eero.simoncelli@nyu.edu

Abstract

We develop a statistical characterization of natural images in the wavelet transform domain. This characterization describes the joint statistics between pairs of subband coefficients at adjacent spatial locations, orientations, and scales. We observe that the raw coefficients are nearly decorrelated, but their magnitudes are highly correlated. A linear magnitude predictor coupled with both multiplicative and additive uncertainties accounts for the joint coefficient statistics of a wide variety of images including photographic images, graphical images, and medical images. In order to directly demonstrate the power of this model, we construct an image coder called EPWIC (Embedded Predictive Wavelet Image Coder), in which subband coefficients are encoded one bitplane at a time using a non-adaptive arithmetic encoder that utilizes probabilities calculated from the model. Bitplanes are ordered using a greedy algorithm that considers the MSE reduction per encoded bit. The decoder uses the statistical model to predict coefficient values based on the bits it has received. The rate-distortion performance of the coder compares favorably with the current best image coders in the literature.

1 Introduction

Many applications in image processing require a statistical prior model. In image compression, the theoretical limits of an algorithm are defined by the prior model used, though this model is often implicitly defined. In this paper, we describe an explicit prior probability model for images in the wavelet transform domain, and test this model by using it in an image compression algorithm. The resulting algorithm is quite flexible, and well-suited for encoding of images that must be retrieved over a variety of communication links.

Orthonormal wavelet pyramids, in which images are decomposed using basis functions localized in spatial position, orientation, and spatial frequency (scale), have proven to be extremely effective for

• RB is supported by NSF Graduate Fellowship GER93-55018. EPS is supported by NSF CAREER grant MIP-9796040, ARO/MURI DAAH04-96-1-0007, and the Sloan Center for Theoretical Neurobiology at NYU.
• Preliminary versions of this work have been published in [2] and [24].

image compression [e.g., 26, 27, 7, 1, 20, 18]. We believe there are several statistical reasons for this success. The most widely known of these is that wavelet transforms are reasonable approximations to the Karhunen-Loève expansion for fractal signals [28], such as natural images [17]. The subbands of an orthonormal wavelet decomposition have a wide range of variances whose sum is equal to that of the original image. If the subbands are encoded with a simple first-order entropy encoder, the minimum coding size of the image representation is sum of the entropies of the subbands. Since entropy is a concave function, the differing variances result in a coding cost significantly less than the first-order entropy of the raw image.

In addition to this redistribution of variance, wavelet transforms produce coefficients with significantly non-Gaussian statistics, and thus have lower entropy than a Gaussian-distributed signal of the same variance. This property has been exploited in compression, noise removal and texture synthesis [e.g., 11, 5, 8, 30, 23]. We discuss it in greater detail in section 2, and provide an explicit model for these statistics.

Finally, wavelet decompositions exhibit joint statistical regularities that have been implicitly utilized in a number of recent image coding algorithms [13, 20, 16, 19, 9, 3, 29]. These regularities are the primary topic of this paper. We discuss them in greater detail in section 3, and provide an explicit statistical model describing the relationships between coefficients of different subbands.

In order to demonstrate the quality of our statistical model, the latter half of the paper describes an embedded predictive wavelet image coder that directly utilizes the model. Section 4 describes the compression algorithm, and the details concerning the implementation of the coder. Finally, section 5 analyzes the performance of the coder, and compares it to several standard coders.

2 First-order Subband Statistics

A number of authors have observed that wavelet subband coefficients have highly non-Gaussian statistics [e.g., 11, 6, 12, 23]. The intuitive explanation for this is that images typically have spatial structure consisting of smooth areas interspersed with occasional edges or other abrupt transitions. The smooth regions lead to near-zero coefficients, and the structures give occasional large-amplitude coefficients.

Histograms[1] for wavelet subbands of several images are plotted in figure 1. Compared to a Gaussian, these densities are more sharply peaked at zero, with more extensive tails. To quantify this, we give the sample kurtosis (fourth moment divided by squared second moment) below each histogram. The estimated kurtoses of all of the subbands are significantly larger than the value of three expected for a Gaussian distribution. These examples were computed for subbands of an orthonormal separable wavelet decomposition (see section 4 for details), but they would be similar for any octave-bandwidth subbands.

These non-Gaussian densities should be contrasted with statistics of frequency-based decompositions which are approximately Gaussian. Since the Gaussian is the maximal-entropy distribution for a given variance, wavelet-based coders are able to achieve higher degrees of compression than frequency-

[1]By considering these as representative of the underlying coefficient densities, we are making an implicit assumption of spatial stationarity.

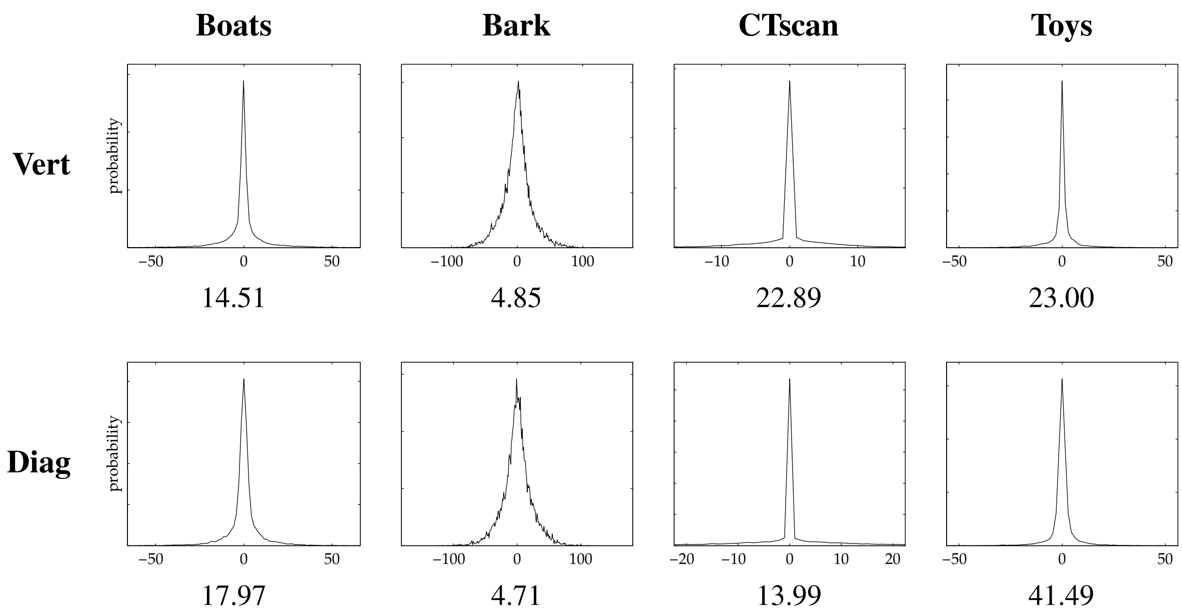

Figure 1: Example wavelet subband coefficient histograms. Subbands correspond to four different images (a landscape, a texture, a medical image, and a synthetic graphics image) and two different orientations (top: vertical, bottom: diagonal), at an intermediate scale. Original images are shown in figure 11. Below each histogram is the sample kurtosis (fourth moment divided by squared variance). A Gaussian density has a kurtosis of 3.

based coders such as JPEG. The non-Gaussianity of wavelet marginals may be taken as an indication that the Wavelet basis is more appropriate for image representation than either pixel or Fourier representations.

These coefficient statistics have been previously modeled [11, 23] using a two-parameter "generalized Laplacian" density function of the form:

$$f_{s,p}(c) = \frac{e^{-|c/s|^p}}{N(s,p)}, \tag{1}$$

where $N(s,p) = 2s\Gamma(1/p)/p$, and $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\,dt$, the Gamma function. The parameters $\{s,p\}$ are directly related to the second and fourth moments. Specifically:

$$\sigma^2 = \frac{s^2\Gamma(\frac{3}{p})}{\Gamma(\frac{1}{p})}, \qquad \kappa = \frac{\Gamma(\frac{1}{p})\Gamma(\frac{5}{p})}{\Gamma^2(\frac{3}{p})}, \tag{2}$$

where σ^2 is the distribution variance, κ is the kurtosis.

We solve for the parameters $\{s,p\}$ by minimizing the relative entropy (also known as the "Kullback-Leibler distance") between a discretized model distribution and the 256-bin coefficient histogram:

$$E(s,p) = \sum_{n=1}^{256} h_n \log_2 \frac{\bar{f}_{s,p}(c_n)}{h_n},$$

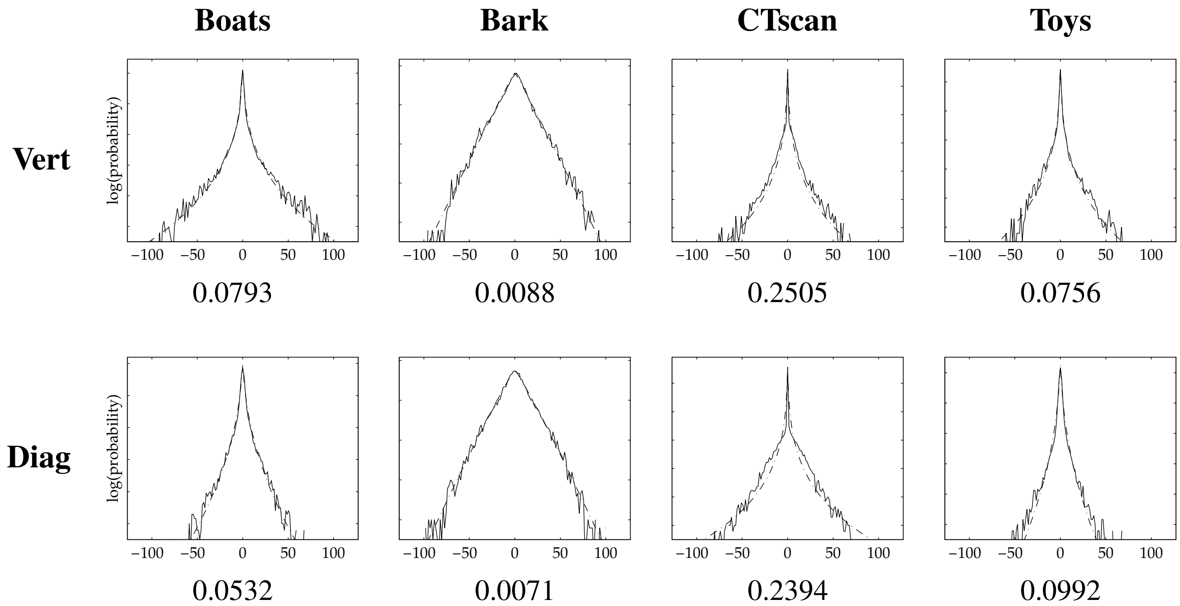

Figure 2: Examples of 256-bin coefficient histograms (solid lines) fitted with the density of equation (1) (dotted lines), plotted in the log domain. Subbands are the same as those in figure 1. Below each graph is the relative entropy (Kullback-Liebler distance) between the histogram and the model of equation (1). The CTscan image was the worst (in terms of relative entropy) in our image set.

where $\bar{f}_{s,p}(c_n)$ is the integral of the density given in equation (1) over the nth histogram bin (centered at value c_n), and h_n is the normalized histogram count (frequency) for histogram bin n. The measure $E(s,p)$ corresponds to the cost (in bits) of encoding the data with an entropy coder that assumes the distribution $f_{s,p}(c)$. For the images in our sample set, σ^2 is approximately proportional to 2^{-l} (where l indicates the pyramid level), and the parameter p is typically in the range $[0.5, 1.0]$, corresponding to a kurtoses in the range $[6, 25.2]$.

We make no claim of optimality for this model: Other authors [e.g., 30] have used alternative density functions to describe these distributions. Nevertheless, the fits are quite good. Figure 2 shows log-domain plots of the histograms of figure 1, together with plots of the fitted density function of equation (1). We have included both the best and worst cases from the set of images in our test set (figure 11). Below each figure is the relative entropy between the histograms and fitted densities.

Figure 3 shows a scatterplot comparing the encoding cost using the model of equation (1), and the encoding cost assuming a Gaussian density vs. the encoding cost assuming accurate knowledge of a 256-bin histogram. The Gaussian examples were computed with the distribution variance matched to the sample variance. Note that the relative entropy of the Laplacian model is less than 0.25 bits/coefficient for our sample images, as compared with the Gaussian density model which often has a relative entropy greater than 1.0 bit/coefficient.

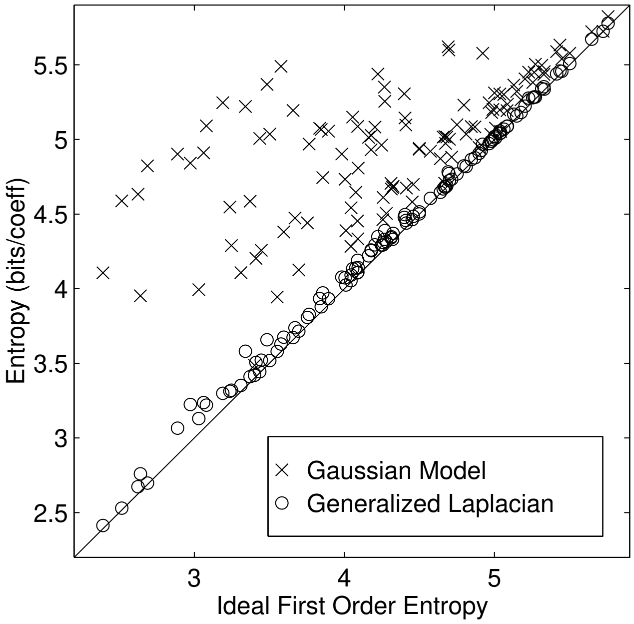

Figure 3: Comparison of encoding costs. Plotted are encoding cost assuming the generalized Laplacian density of equation (1) (O's), and the encoding cost assuming a Gaussian density (X's), versus the encoding cost using a 256-bin histogram. Points are plotted for 9 bands (3 scales, 3 orientations) of the 13 images in the sample set of figure 11.

3 Joint Subband Statistics

As mentioned in the introduction, the coefficients of wavelet subbands are nearly decorrelated. Nevertheless, it is clear from casual inspection that wavelet coefficients are *not* statistically independent. Figure 4 shows the magnitudes of wavelet coefficients in a four-level pyramid decomposition. The large-magnitude coefficients tend to occur at the same relative spatial locations in subbands at adjacent scales [2].

Such dependencies are utilized implicitly in a number of recent image compression schemes. Shapiro [20] constructed the Embedded Zerotree Wavelet (EZW) coder to exploit the fact that a coefficient is likely to have small magnitude if the coefficients at coarser scales have small magnitudes. The Zerotree technique encodes entire trees of zeros with a single symbol, thus capturing a portion of the conditional distribution of a coefficient given its parent, grandparent, etc.

Several authors [14, 13, 16] have used vectorized lookup tables to predict blocks of fine coefficients from blocks of coarse coefficients. Schwartz et. al. [19] used adaptive entropy coding to capture conditional statistics of coefficients based on the most significant bits of each of the eight spatial neighbors and the coefficient at a coarser scale. Chrysafis and Ortega [3] switch between multiple probability models depending on values of neighboring coefficients. Said and Pearlman [18] use a predictive scheme to give high-quality zerotree coding results, and Wu and Chen [29] have extended the EZW coder to use local coefficient "contexts".

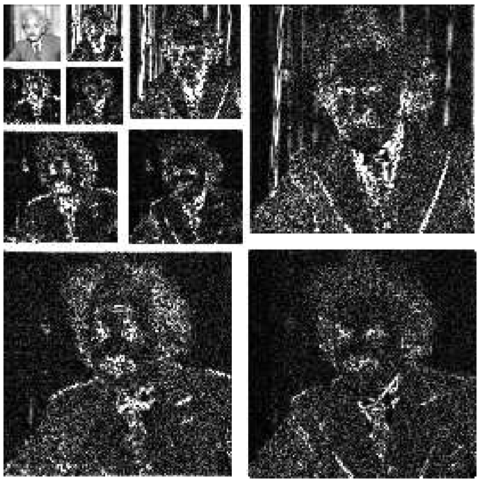

Figure 4: Coefficient magnitudes of a wavelet decomposition. Shown are absolute values of subband coefficients at three scales, and three orientations of a separable wavelet decomposition of the Einstein image. Also shown is the lowpass residual subband (upper left). Note that high-magnitude coefficients of the subbands tend to be located in the same (relative) spatial positions.

3.1 Joint Magnitude Statistics

We wish to explicitly examine and utilize the statistical relationship between wavelet coefficient magnitudes. Consider two coefficients representing information at adjacent scales, but the same orientation (e.g, horizontal) and spatial location. As in the previous section, we will assume spatial stationarity, which allows us to consider the joint histogram of this pair of coefficients as representative of the underlying statistics. Figure 5A shows the log-domain conditional histogram $\mathcal{H}\left(\log_2(C)|\log_2(P)\right)$, where P is the magnitude of the coarse-scale ("parent") coefficient and C is the magnitude of the finer-scale ("child") coefficient.

Observe that the right side of the distribution is unimodal and concentrated along a unit-slope line. This suggests that in this region, the conditional expectation, $\mathcal{E}(C|P)$, is approximately proportional to P. Furthermore, vertical cross sections (i.e., conditional histogram for a fixed value of P) have approximately the same shape for different values of P. Finally, the left side of the distribution is concentrated about a horizontal line, suggesting that C is independent of P in this region. We suspect these low-amplitude coefficients are dominated by quantization and other noise sources.

The intuition behind this joint probability relationship is that typical localized image structures such as edges tend to have substantial power at a range of scales and orientations. Thus, near such an image feature, the coefficients at neighboring scales will tend to have large magnitudes. A simple simulation confirms this intuition. We computed the joint density of horizontal wavelet coefficient magnitudes at two adjacent scales for a 64 $\times$ 64 pixel image of a disk. The disk radius was drawn uniformly from the interval $[8, 24]$ pixels, and its position was randomized within ± 8 pixels of the image center. The disk amplitude was drawn from a generalized Laplacian density, with parameters $s = 1, p = 0.5$. We also added a small amount (SNR = 30 dB) of uniformly distributed white noise to the image. We collected statistics over 400 such images. The resulting distribution, plotted in figure 5B, is qualitatively similar to that of figure 5A.

The form of the histogram shown in figure 5A is surprisingly robust across a wide range of images.

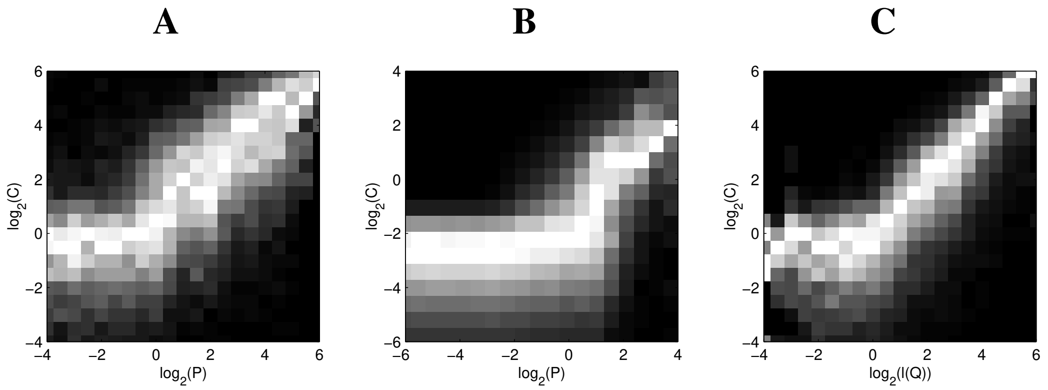

Figure 5: Conditional histograms for a fine scale horizontal coefficient. Brightness corresponds to probability, except that each column has been independently rescaled to fill the full range of display intensities. **A:** Conditioned on the Parent (same location and orientation, coarser scale) coefficient. Data are for the Boats image. **B:** Same as **A**, but data are for a synthetic image containing non-overlapping disks of randomized spatial position and size and additive noise (see text). **C** Conditioned on a linear combination of neighboring coefficient magnitudes. Data are for the same subband of the Boats image.

Furthermore, the qualitative form of these statistical relationships also holds for pairs of coefficients at adjacent spatial locations (which we call "siblings"), adjacent orientations ("cousins"), and adjacent orientations at a coarser scale ("aunts"). This set of potential conditioning coefficients (we refer to these as "neighbors") is illustrated in figure 6.

3.2 Linear Magnitude Predictor

Given the linear relationship between the magnitudes of large-amplitude coefficients, and the difficulty of characterizing the full multi-dimensional density, we chose to examine a linear predictor for coefficient magnitude:

$$l(\vec{Q}) \equiv \vec{w} \cdot \vec{Q} = \sum_k w_k Q_k, \qquad (3)$$

where the coefficient magnitude set $\{Q_k\}$ corresponds to a subset of the potential conditioning neighbors, as depicted in figure 6. The weights w_k of the coefficients in the linear estimator are chosen to minimize the expected squared error of the estimator. That is:

$$\vec{w} = \mathcal{E}(\vec{Q}\vec{Q}^T)^{-1} \cdot \mathcal{E}(C \cdot \vec{Q}), \qquad (4)$$

where $\mathcal{E}(\cdot)$ indicates the expected value of a random variable, C corresponds to the coefficient magnitude being estimated, and $\vec{Q}$ is a random vector containing the magnitudes of the conditioning neighbors of C.

Figure 5C shows a conditional histogram, $\mathcal{H}\left(\log_2(C)|\log_2(l(\vec{Q}))\right)$ based on magnitudes of eight adjacent coefficients in the same subband, two cousin coefficients, and one parent coefficient (interpolated to the correct position using bilinear interpolation). Note that the distribution has a similar appearance to the single-parent distribution of figure 5A. But the linear region is extended, and the conditional variance is greatly reduced.

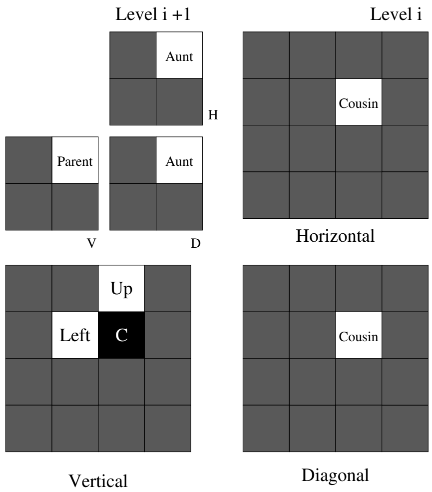

Figure 6: Subset of wavelet coefficients surrounding a given coefficient (C), that are potentially suitable for conditioning.

Child subband	Last neighbor included in predictor					
Horizontal	Left	Up	Parent	DiagCousin	LeftLeft	DiagAunt
	0.3322	0.4308	0.4635	0.4804	0.4903	0.4939
Vertical	Up	Left	Parent	DiagCousin	UpUp	DiagAunt
	0.3513	0.4356	0.4675	0.4865	0.4929	0.4987
Diagonal	Up	Left	Parent	Horiz Cousin	Vert Cousin	Left Left
	0.2175	0.2792	0.3134	0.3235	0.3294	0.3356

Table 1: Cumulative mutual information between coefficient magnitude C and a linear estimator $l(\vec{Q})$ composed of a subset of neighbors. Each entry gives the mutual information for a subset containing the neighbors indicated at the top of that column and all columns to the left. Notice that the local neighbors within the subband (Left and Up), the Parent, and the Cousins contribute most to the mutual information. Values are averaged over two scales (levels 1 and 2) of three training images (Lena, Boats, Baboon).

In order to determine which coefficients to include in the conditioning set $\{Q_k\}$, we calculated the mutual information between C and $l(\vec{Q})$ for a variety of choices of interband and intraband coefficients in $\{Q_k\}$. The mutual information gives the theoretical coding gain (in bits per coefficient) obtained when encoding C using the conditional histogram $\mathcal{H}\left(C|l(\vec{Q})\right)$ (i.e., assuming $l(\vec{Q})$ is known to the receiver) compared with encoding C using only the marginal histogram $\mathcal{H}(C)$. Rather than exhaustively explore all possible neighbor subsets, we used a greedy algorithm to choose conditioning neighbors. Table 1 shows the greedy optimal neighbor subset for the three oriented subbands. Using this analysis, and imposing causality (assuming a standard scanline ordering of the bits), we decided to include neighbors corresponding to the first four table columns when coding the horizontal and vertical bands, and the first five columns for the diagonal bands.

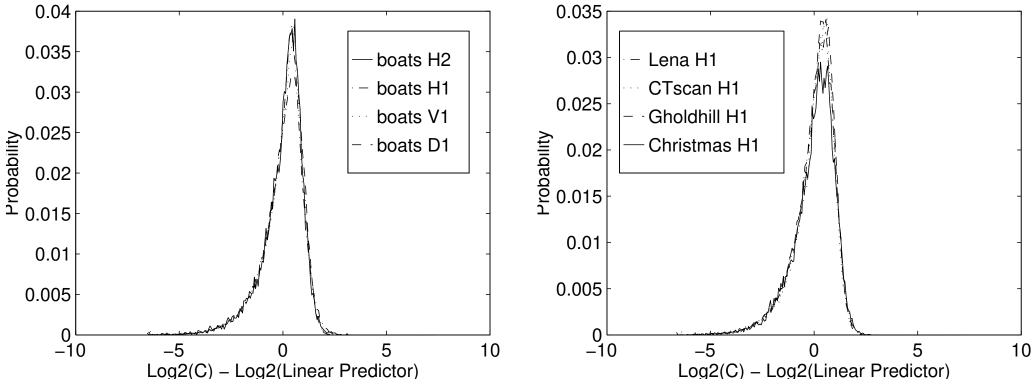

Figure 7: Comparison of conditional distributions in the log domain of different subbands and images. Distributions were normalized (in the log domain) to have mean 0 and variance 1. **Left:** Comparison of distributions for different subbands of the image 'boats'. **Right:** Comparison of distributions for different images (Lena, Goldhill, CTscan, Christmas).

3.3 Conditional Probability Model

Based on the linear magnitude predictor described in the previous section, we wish to characterize the joint densities of the coefficients. We were surprised to observe that the conditional distribution in the log domain, when normalized for mean and variance, is highly consistent across subbands of an image, and even across a wide range of images. Figure 7 shows a comparison of these conditional distributions for four subbands from the "Boats" image, and also a comparison of a single band across four different images. We included only the right portion of each conditional histogram (i.e., the region in which C is proportional to $l(\vec{Q})$).

The fact that the conditional histograms seem to have a constant shape that shifts linearly with the predictor in the log domain suggests a model of multiplicative uncertainty. In particular, we use the following generative model for the conditional density:

$$C' = M \cdot l(\vec{Q}) + N, \tag{5}$$

where C' is the signed coefficient (i.e., $C = |C'|$), $l(\vec{Q})$ is the linear magnitude predictor described previously, and M and N are two mutually independent zero-mean random variables. Note that C' will be uncorrelated with each of the conditioning coefficients, Q_k.

To model the distribution of M, we constructed a function $G(\cdot)$ to represent the cumulative of the conditional density in the log domain, by averaging the mean- and variance-normalized conditional histograms of three training images (Lena, Boats, Baboon), at two scales (levels 2 and 3) and all three orientations. We use this is as a parameterized model for the conditional cumulative:

$$\mathcal{P}\left(C < c \mid l(\vec{Q}) = p\right) \approx G\left(\frac{\log_2(c) - \log_2(p)}{\alpha}\right), \tag{6}$$

where α parameterizes the width for for each subband. We assume N is independent of M, and Gaussian-distributed.

Given these distributional assumptions, the model described by equation (5) is characterized by the linear weights $\{w_k\}$, the variance α^2 of M (in the log domain), and the variance σ^2 of N. The linear weights are chosen via equation (4) to be least-squares optimal. The variance parameters are then computed by minimizing the relative entropy between the joint model density and the joint histogram. Figure 8 shows comparisons of joint histograms of the second-level horizontal subband of four different images, with plots of the fitted density function generated by equation (5).

An entropy calculation shows the value and quality of the model. Figure 9 shows a scatterplot comparing encoding cost based on the joint probability model of equation (5) vs. the encoding cost assuming accurate knowledge of a 256 $\times$ 256-bin histogram. Also included is a comparison to the first-order histogram entropies. The conditional model falls short of the ideal by less than 0.7 bit. In these situations, the conditional histogram (from which the ideal values in the table are computed) for large-magnitude predictors is sparse and has high variance. The predictive models, however, are based on *smooth* high-variance densities. Thus, the ideal values are deceptively low due to detailed knowledge of the coefficient values for this specific subband. Nevertheless, the linear-predictive model is substantially better than the first-order model, consistent with the mutual entropy estimates of table 1.

Finally, we should consider the signs of the coefficients. The lowpass bands contain almost entirely positive coefficients: over our sample set of 13 images, the percentage of negative coefficients is 2.4%. In addition, we model the lowpass coefficient distribution as a uniform density. For the bandpass subbands, the probability of positive and negative coefficients is equal. They are, however, not spatially independent. In the horizontal bands, for example, the probability of the "Up" neighbor having the same sign is 36%. There are also more complex relationships between sign bits in neighboring subbands, but we will not attempt to characterize those in this paper.

4 Implementation of a Progressive Image Coder

In this section we describe the implementation of our Embedded Predictive Wavelet Image Coder (EPWIC), based on the conditional probability model developed in Section 3.3. Our implementation is simple and reasonably efficient, but comes quite close to the theoretical entropy associated with the probability model. In addition, it compares favorably with the current best coders in the image processing literature.

4.1 Separable Wavelet Decomposition

For our coder, we utilize a recursive pyramid decomposition based on separable Quadrature Mirror Filters (QMFs). The one-dimensional lowpass and highpass kernels are linear-phase (i.e., symmetric) 9-tap filters, and are optimized to nearly satisfy the one-dimensional orthonormal system diagram shown in figure 10, as described in [22].

The two one-dimensional kernels, $L(\omega)$ and $W(\omega)$, are applied separably along the axes of the image sampling lattice in order to generate a single level of a wavelet pyramid. This consists of vertical, horizontal and diagonal subbands, and a lowpass subband. Subsequent pyramid levels (i.e., subbands at different scales) are created by applying this 4-band splitting procedure recursively to the lowpass

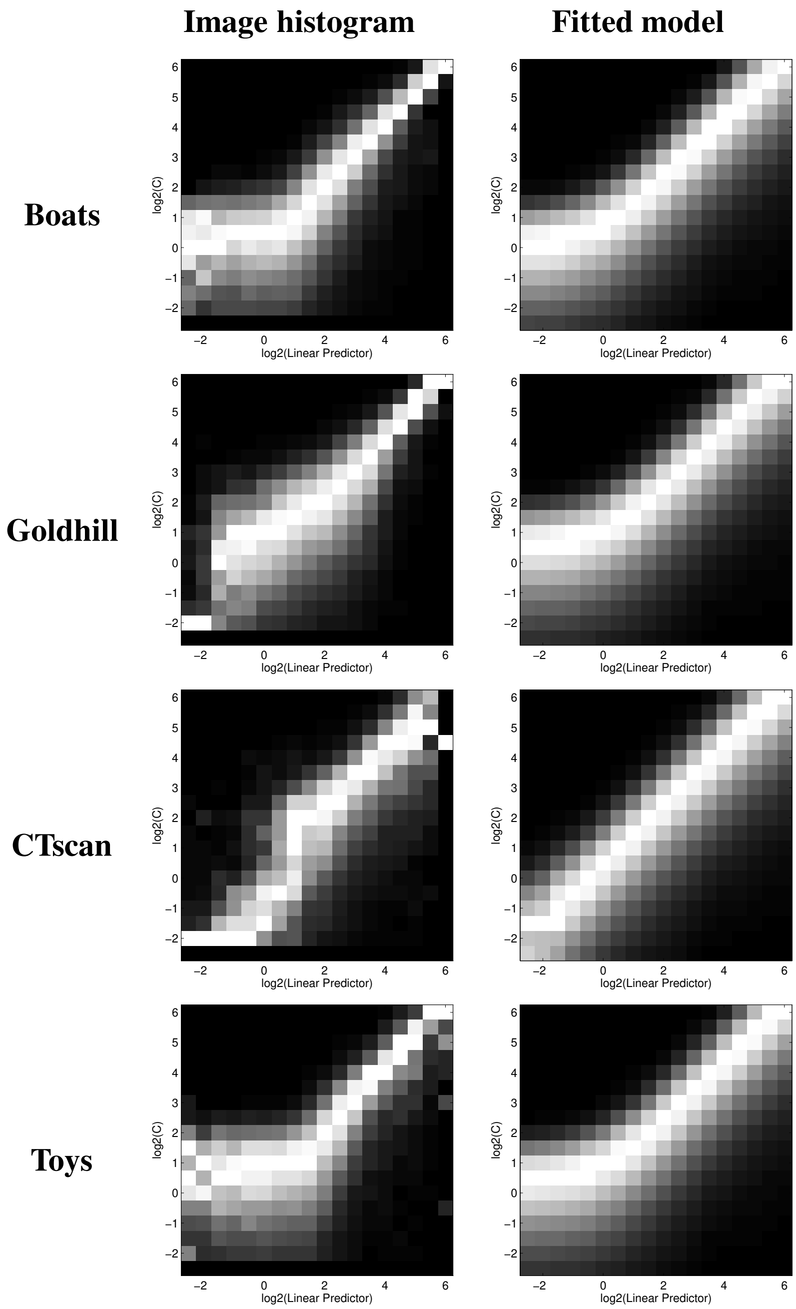

Figure 8: Left: Examples of log-domain conditional histograms for the second-level horizontal subband of different images, conditioned on an optimal linear combination of coefficient magnitudes from adjacent spatial positions, orientations, and scales. **Right**: Model of equation (5) fitted to the conditional histograms in the left column. Intensity corresponds to probability, except that each column has been independently rescaled to fill the full range of intensities.

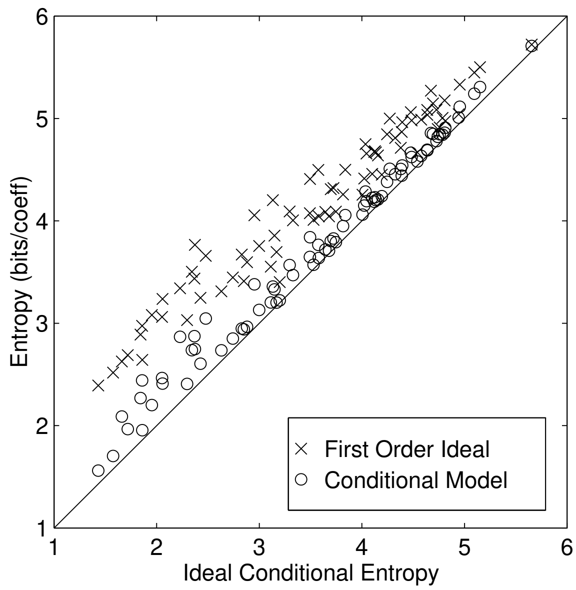

Figure 9: Comparison of encoding cost using the conditional probability model of equation (5), and the encoding cost using the first-order model (i.e., equation (1)) as a function of the encoding cost using a 256×256-bin joint histogram. Points are plotted for 6 bands (2 scales, 3 orientations) of the 13 images in our sample set.

subband. Convolution boundaries are handled by symmetric reflection of the image about the edge pixels, as described in [21]. Reconstruction PSNR from 5-level pyramids on the images in our sample set is typically over 63 dB.

We denote the basis functions in the separable wavelet transform as $w_{\{o,s,n,m\}}(x, y)$, where $o \in \{1, 2, 3\}$ indicates the orientation (vertical, horizontal, diagonal, respectively), s indicates the pyramid level (scale), and (n, m) indicates the spatial location of the basis function. These basis functions are normalized to have unity L_2-norm. The wavelet representation consists of the set of coefficients $\{c(o, s, n, m)\}$ corresponding to each of the basis functions.

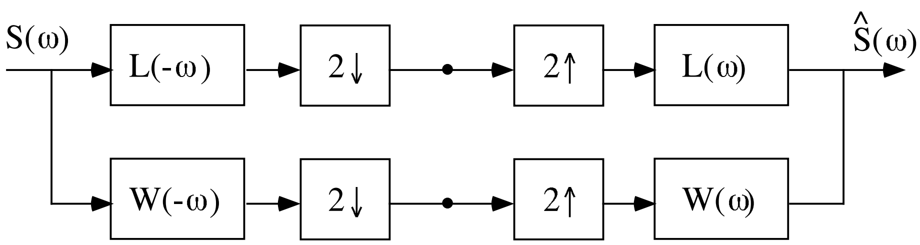

Figure 10: One-dimensional system diagram for discrete dyadic wavelet decomposition. The two filters are related via modulation, time-reversal, and a one-sample shift: $W(\omega) = e^{j\omega} L(-\omega + \pi)$. The system response is $|W(\omega)|^2 + |L(\omega)|^2$.

4.2 Coefficient Bitplane Encoding

In order to have maximal control over the ordering of image information, we consider each subband coefficient to be quantized to a 16-bit binary integer (including sign bit). That is:

$$c(o, s, n, m) = a(s, o)\rho(o, s, n, m) \sum_k 2^k b(o, s, n, m, k),$$

where $a(s, o)$ is a scalar multiplier for each subband, $\rho(o, s, n, m)$ is the sign (± 1), and $b(o, s, n, m, k)$ corresponds to the kth bit of the coefficient $c(o, s, n, m)$.

The wavelet decomposition describes the image $I(x, y)$ as a linear combination of the basis functions:

$$\begin{aligned} I(x, y) &= \sum_{o,s,n,m} a(s, o)\, \rho(o, s, n, m) \sum_k 2^k\, b(o, s, n, m, k)\, w_{\{o,s,n,m\}}(x, y) \\ &= \sum_{o,s,n,m,k} b(o, s, n, m, k) \left[a(s, o)\rho(o, s, n, m) 2^k w_{\{o,s,n,m\}}(x, y) \right]. \end{aligned}$$

The second line suggests that we may view this as a type of representation in which the coefficients are restricted to the set $\{0, 1\}$ (i.e., they are single-bit quantities). The basis functions of this representation are $w'(o, f, n, m, k) = a(s, o)\rho(o, s, n, m)2^k w_{\{o,s,n,m\}}(x, y)$, which are related to each other by translation, dilation, Fourier modulation, negation, and *scalar multiplication by powers of two*.

Progressive transmission of an image requires us to choose an ordering of the coefficient bits. In order to keep the complexity of the coder down, we assume that all bits at a given significance level k of a subband will be sent consecutively, in raster order. We refer to this collection of bits as a *bitplane*. We also assume that the bitplanes of a given subband will be sent in order from most to least significant. Since most of the coefficient values are close to zero, the sign bit of each coefficient is sent only when needed, immediately after the first non-zero magnitude bit, as in [19].

In general, the ordering of bitplanes across subbands should take into account both the encoded size of the bitplane and the improvement in decoded image quality resulting from the incorporation of that bitplane. We use a greedy algorithm (which we refer to as a "bang-for-the-buck" algorithm), in which we select the bitplane with the maximum ratio of MSE reduction per encoded bit. That is, at each point we choose the bitplane which produces the most steeply descending rate-distortion curve.

The MSE reduction is simple to calculate. Since the wavelet basis is orthonormal, the squared error of an image is the sum of the squared errors of each of the coefficients in the pyramid. Therefore the improvement in MSE for a bitplane is simply the number of nonzero bits in the bitplane multiplied by the squared magnitude represented by the bits in that bitplane.

The encoded size of a bitplane is measured as the smallest of three possible representations: raw encoding, run-length encoding using block sizes, b, of either 8 or 16 bits, and non-adaptive arithmetic encoding based on our probability model (described in the next section).

Our run-length encoding algorithm encodes the length of strings of consecutive blocks of zeros in the data stream. If a string of zero blocks is encountered in the input stream, the first bit of the output block is set to 0 and the rest of the block contains the (binary-encoded) number of zero blocks (up to $2^{(b-1)}$). Otherwise the first bit of the output block is set to 1, and the remainder of the block contains $b - 1$ raw bits from the input stream.

The arithmetic encoder is similar to the algorithm found in [pp. 910-915, 15], which encodes a data stream using a probability distribution that is adaptively computed and stored in a histogram. Instead of computing such a histogram, our encoder uses the distribution specified by our statistical model. Since the "symbols" of our input stream are single bits, the probability that a bit is non-zero is all that is needed to construct the arithmetic code.

4.3 Calculation of Bit Probabilities

Our encoding technique makes direct use of the model joint probability density described earlier. In particular, both encoder and receiver must use this distribution to compute the conditional mean estimate for each coefficient, given the bits that have been sent/received thus far. In addition, as described above, the arithmetic coder and decoder must know the probability that a given bit will be non-zero.

Consider the arithmetic encoding of the kth bit of a particular subband coefficient, C_k. The encoder must calculate the probability that the bit is nonzero, given the set of all coefficient bits that have already been received. The set of bits received constrain the magnitude of the coefficient of interest (C), and the magnitudes of each of the conditioning coefficients $\{Q_n\}$ to lie in particular ranges:

$$\begin{aligned} C &\in [l_0, h_0] \\ Q_n &\in [l_n, h_n], \qquad n = 1, 2, \ldots, N. \end{aligned} \tag{7}$$

The probability that we wish to calculate is:

$$\begin{aligned} &\mathcal{P}\left(C_k = 1 \mid \text{bits received thus far}\right) \\ &\qquad = \frac{\mathcal{P}\left(m_0 < C < h_0 \mid l_n < q_n < h_n,\ \forall n\right)}{\mathcal{P}\left(l_0 < C < h_0 \mid l_n < q_n < h_n,\ \forall n\right)} \\ &\qquad = \frac{\int dp\, \mathcal{P}\left(m_0 < C < h_0 \mid l(\vec{Q}) = p\right)\ \mathcal{P}\left(l(\vec{Q}) = p \mid l_n < q_n < h_n,\ \forall n\right)}{\int dp\, \mathcal{P}\left(l_0 < C < h_0 \mid l(\vec{Q}) = p\right)\ \mathcal{P}\left(l(\vec{Q}) = p \mid l_n < q_n < h_n,\ \forall n\right)} \end{aligned} \tag{8}$$

where $m_0 = (l_0 + h_0)/2$. The last expression uses the assumption of our joint probability model: given $l(\vec{Q})$, the values of $\vec{Q}$ do not provide any more information about C, and therefore:

$$\mathcal{P}\left(l_0 < C < h_0 \mid l(\vec{Q}) = p, l_n < q_n < h_n,\ \forall n\right) \approx \mathcal{P}\left(l_0 < C < h_0 \mid l(\vec{Q}) = p\right).$$

In order to avoid the computationally expensive integration over p, we use two approximations in the conditional probability of C that allow us to perform a simple one-dimensional calculation. First, we assume that the noise N plays a small role in this function. In this case, we may normalize C by its standard deviation, $\sqrt{l^2(\vec{Q}) + \sigma^2}$, to get a random variable that is characterized by the model cumulative distribution of equation (6). Specifically, we use the approximation:

$$\mathcal{P}\left(l_0 < C < h_0 \mid l(\vec{Q}) = p\right) \approx G\left(\frac{\log_2(h_0) - \log_2(\sqrt{p^2 + \sigma^2})}{\alpha}\right) - G\left(\frac{\log_2(l_0) - \log_2(\sqrt{p^2 + \sigma^2})}{\alpha}\right),$$

where σ is the standard deviation of N.

Second, we eliminate the dependence on p by replacing it with its conditional mean, and correct for this by broadening the conditional distribution. That is:

$$\int dp\, \mathcal{P}\left(m_0 < C < h_0 | l(\vec{Q}) = p\right) W(p)$$
$$\approx \left[G\left(\tfrac{\log_2(h_0) - \log_2(\sqrt{\hat{p}^2+\sigma^2})}{\alpha'}\right) - G\left(\tfrac{\log_2(m_0) - \log_2(\sqrt{\hat{p}^2+\sigma^2})}{\alpha'}\right)\right] \int dp\, W(p),$$

where

$$W(p) = \mathcal{P}\left(l(\vec{Q}) = p \,|\, l_n < q_n < h_n,\ \forall n\right),$$

$\hat{p}$ is the conditional mean of the predictor:

$$\hat{p} \;=\; \sum_n w_n \int_{l_n}^{h_n} dq\; f_{s_n, p_n}(q)\; q,$$

and the scalar α' depends on both the width of the log-domain conditional density (as before) and on $W(p)$, and serves to broaden the approximated density. The value of α' can be determined for each bitplane of a subband through gradient descent to minimize the arithmetic encoded size of the bitplane.

Finally, substituting this approximation into the original expression, and eliminating common factors, gives:

$$\mathcal{P}\left(C_k = 1 | \text{bits received thus far}\right)$$
$$\approx \frac{G\left(\tfrac{\log_2(h_0) - \log_2(\sqrt{\hat{p}^2+\sigma^2})}{\alpha'}\right) - G\left(\tfrac{\log_2(m_0) - \log_2(\sqrt{\hat{p}^2+\sigma^2})}{\alpha'}\right)}{G\left(\tfrac{\log_2(h_0) - \log_2(\sqrt{\hat{p}^2+\sigma^2})}{\alpha'}\right) - G\left(\tfrac{\log_2(l_0) - \log_2(\sqrt{\hat{p}^2+\sigma^2})}{\alpha'}\right)}. \tag{9}$$

This is the expression used for the calculation of bit probabilities in the coder.

4.4 Summary of EPWIC Compression Algorithm

To encode an image $I(x, y)$:

1. Choose a goal MSE (we typically use the variance associated with the quantization of the original image).
2. Calculate the coefficients $c(o, s, n, m)$ of the wavelet decomposition.
3. For each subband in the decomposition, quantize coefficients (to 16 bits) and retain the quantization binsize, $a(s, o)$.
4. Characterize the statistics of the subband at each orientation and scale (o, s):
   (a) Calculate the parameters s and p that minimize the relative entropy between the histogram of the coefficients and the density of equation (1).

(b) Calculate least-squares optimal weights $\vec{w}$ of the linear predictor $l(\vec{Q})$ (equation (4)).

(c) Calculate σ^2 (the variance of N) that minimize the relative entropy between the joint coefficient histogram and the probability density described by equation (5).

5. Transmit an EPWIC identification tag (16 bits), the width and height of the image (16 bits each), and the number of levels in the pyramid (8 bits).

6. While the decoded MSE is greater than the goal MSE:

   (a) Determine which of the set of candidate bitplanes (i.e., the most significant remaining bitplanes of each subband) should be encoded next, by comparing the "bang-for-the-buck" for each candidate bitplane:

   - Calculate the MSE reduction associated with the content of the bitplane.
   - Arithmetic encode the bitplane. Update the conditional mean estimates of any neighbors, $\hat{Q}_n$, that have changed , and update the linear predictor $\hat{p}$. Calculate α', to minimize the relative entropy of the model density compared with the coefficient histogram. For each bit in the bitplane, calculate the probability of the bit being nonzero (equation (9)), and construct a code stream using the arithmetic coder.
   - Run-length encode the bitplane, using both the 8-bit and 16-bit run-length encoding methods.
   - Choose the minimum of the ratio of MSE reduction to coding size for the three coding methods.

   (b) Transmit a tag identifying the subband to which the bitplane belongs, and indicating the encoding method (8 bits: 2 bits indicate the method, and 6 indicate the subband).

   (c) If this is the first encoded bitplane of this subband, transmit the quantization binsize $a(s, o)$ (16 bits, representing the interval [0, 29]), the s value (8 bits, representing [0.5, 128.5]), the p value (8 bits, representing $[0.2, 2]$), and the value of σ (8 bits, representing $[2^{-5}, 2^5]$).

   (d) If arithmetic-encoding the bitplane:

   - If this is the first arithmetic-encoded bitplane of this subband, transmit $\{w_k\}$ (8 bits each, representing $[0, 1]$).
   - Transmit α' (8 bits, representing $[0.02, 5.15]$).

   (e) Transmit the encoded data.

5 Results

In order to demonstrate the performance of EPWIC, we encoded the set of 13 images shown in figure 11. Each image was decomposed into a discrete wavelet pyramid containing 5 levels. For comparison purposes, we considered three other image coders:

1. EPWIC-1: we implemented a progressive encoder utilizing the marginal (generalized Laplacian) density of equation (1) as a model of the first order distribution. The coder is otherwise similar to EPWIC-2, in that it uses the same greedy algorithm for ordering of bitplanes, and uses the same arithmetic coding scheme to encode bitplanes.

Figure 11: Full set of images used in our experiments. Left to right, top to bottom: Abdomen, Baboon, Bark, Boats, Christmas, CTscan Flowers, Frog, Goldhill, Lena, Mountain, Paolina, and Toys. Note that Abodomen and CTscan are medical images, and Toys is a synthetic (computer graphics) image. Images are 8-bit/pixel, and of size 512×512, or 480×480 (except for the Abdomen, which is 352×352).

2. JPEG: we used version 5b of CJPEG, a standard JPEG image coder from the Independent JPEG Group.

3. EZW[2]: as described in [20].

Table 2 lists the PSNR values for EPWIC-2, EPWIC-1, and EZW for six of the images. Note that EPWIC-1 surpasses EZW for many images at intermediate compression levels, and EPWIC-2 surpasses EZW at nearly all compression levels. Figure 12 summarizes these results, by showing the PSNR of each coder (relative to EPWIC-1), averaged over the 13 images in our set. EPWIC-1 outperforms EZW for most compression ratios by about 0.3dB, and EPWIC-2 outperforms EZW by 0.5dB at 1Kbyte, and nearly 1.5dB at 16Kbytes and above.

Also shown in figure 12 is the encoding size (relative to that of EPWIC-1) as a function of target SNR. This gives a sense of how long one would wait during a progressive transmission for a result of a given quality. For example, EZW has a transmission time roughly 30% higher than EPWIC-2 for an image quality of 25dB.

In figure 13, an EPWIC-2 progressive transmission series is given for the Boats image. Wavelet aliasing artifacts are quite noticeable in the early stages of the transmission: these are a consequence of using a critically sampled subband representation. Notice that at 16Kbytes (compared with an original image size of 256Kbytes), the reconstructed image is remarkably close to the original.

[2]We thank the David Sarnoff Research Center for their assistance in the EZW comparisons.

Image	**Coder**	**Bits/Pixel**								
		0.008	0.0016	0.031	0.063	0.125	0.25	0.5	1.0	2.0
Abdomen	EPWIC-2	28.46	30.73	33.81	37.14	40.87	44.28	47.86	52.50	59.62
	EPWIC-1	28.17	30.96	33.82	36.75	40.05	43.34	46.95	51.59	58.16
	EZW	25.95	28.02	30.62	33.69	37.12	40.30	43.74	48.10	55.52
	JPEG	NA	NA	17.87	29.01	35.77	40.08	43.47	47.38	51.94
Boats	EPWIC-2	21.25	23.07	24.99	27.09	29.55	32.60	36.65	41.17	46.63
	EPWIC-1	21.06	22.88	24.66	26.74	28.90	31.57	35.35	39.70	44.68
	EZW	21.34	22.83	24.81	26.86	28.87	31.69	35.58	40.00	45.82
	JPEG	NA	NA	18.29	21.83	27.76	30.88	34.63	39.10	43.54
Christmas	EPWIC-2	18.67	20.00	21.20	22.51	24.10	26.86	31.17	37.16	46.48
	EPWIC-1	18.41	19.87	20.99	22.25	23.65	25.97	29.36	34.71	42.30
	EZW	18.45	19.83	21.11	22.16	23.80	26.53	30.18	36.61	43.88
	JPEG	NA	NA	17.26	19.92	23.17	25.09	27.84	32.78	40.83
Lena	EPWIC-2	21.12	23.16	25.49	27.77	30.67	33.48	36.79	40.10	44.98
	EPWIC-1	20.72	23.01	25.21	27.49	30.03	32.72	35.77	39.09	44.03
	EZW	21.03	22.99	25.01	27.46	30.26	33.32	36.46	39.79	44.64
	JPEG	NA	NA	17.95	21.92	28.24	31.42	34.84	37.95	41.62
Mountain	EPWIC-2	15.22	16.16	17.06	17.91	18.81	20.09	22.14	25.52	32.24
	EPWIC-1	15.17	16.16	17.03	17.83	18.71	19.90	21.90	25.15	31.30
	EZW	14.74	15.59	16.49	17.11	18.15	19.12	21.33	24.14	30.97
	JPEG	NA	NA	NA	14.72	17.74	18.82	20.32	22.52	27.27
Toys	EPWIC-2	21.32	22.81	24.74	27.22	30.20	33.94	38.05	42.23	47.08
	EPWIC-1	20.96	22.68	24.28	26.63	29.35	32.64	36.79	40.75	45.48
	EZW	21.30	22.84	24.47	26.44	29.27	32.79	36.97	41.00	45.85
	JPEG	NA	NA	19.34	23.07	27.72	32.28	36.71	40.85	44.04
WorstCase	EPWIC-2	25.93	26.17	26.73	27.27	28.16	29.24	31.50	34.39	40.20
(figure 14)	EPWIC-1	26.58	29.22	31.13	33.17	35.51	38.28	41.55	45.56	51.58
	EZW	26.99	28.93	30.86	32.57	34.58	37.29	40.50	44.28	49.68
	JPEG	NA	NA	18.33	25.22	29.65	32.71	35.97	39.02	42.84

Table 2: PSNR values $10 \cdot \log_{10}(255^2/\mathrm{MSE})$ at different compression ratios for EPWIC-2, EPWIC-1, EZW, and JPEG. Images are shown in figure 11.

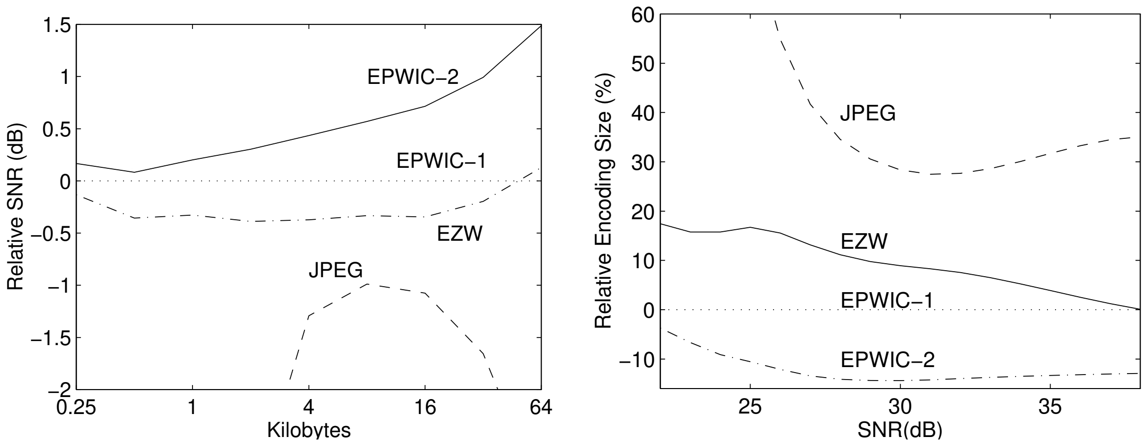

Figure 12: Relative rate-distortion tradeoff for four image coders (JPEG, EZW, EPWIC-1, and EPWIC-2). **Left**: PSNR values (in dB), relative to EPWIC-1 (dotted horizontal line), as a function of the number of encoded bytes. **Right**: Number of bytes necessary to achieve a given PSNR, relative to EPWIC-1 (dotted horizontal line). All curves are averages over the set of 13 images in our collection (shown in figure 11).

Finally, we note that improved compression that EPWIC-2 shows on the images of our test set necessarily means that there must be *some* image for which EPWIC-2 compression will be worse. We synthesized such an image, that is "unlikely" according to the conditional model, but fits the marginal model of EPWIC-1. We began with a set of marginal probability model parameters $\{s, p\}$ from the Lena image. We then filled the subbands of a 4-level pyramid with coefficients drawn independently from the density of equation (1). The resulting image is shown in figure 14, and is qualitatively similar to images that have been generated in [6]. The encoding results for this image are included in table 2. The performance of the coders is quite different than on the sample images: EPWIC-1 encodes this image best, followed by EZW, and EPWIC-2 is worst.

6 Conclusion

We have presented a conditional statistical model for images based on a linear combination of the magnitudes of neighboring coefficients in a wavelet decomposition. The model characterizes the magnitude statistics of a wide variety of images, and provides a useful framework for understanding the compression capabilities of other coders. We have demonstrated the power of the model by using it explicitly in an image coder implementation. The compression results are quite good, especially given the simplicity of the encoding scheme and the fact that we did not utilize the statistical properties of the signs of the coefficients.

Exploitation of sign statistics could yield significant improvements in compression. In particular, the current coder does not make predictions of coefficients before receiving the sign bits. A model that allowed prediction of sign bits from causal neighbors (including those at coarser scales), would allow the coder to fabricate image detail early in a progressive transmission sequence. This type of prediction would also allow the creation of synthetic images with statistics matched to a given sample

Figure 13: A progressive series of the Boats image, encoded using EPWIC-2. Data transmission amounts are given above each image.

image.

There are a number of small improvements that could be made in the implementation of EPWIC. The L_1-norm combination of neighboring magnitudes (described in equation (3)) could be replaced with an L_p-norm predictor, with p chosen to optimize the coding gain. Our preliminary examination of this possibility suggests that our current choice of $p = 1$ is nearly optimal. In addition, the choice of wavelet decomposition in EPWIC is somewhat arbitrary and could be altered. We have not systematically studied this issue, but we expect the coding improvements to be small. Finally, the overhead associated with the model parameters could be reduced. The parameter σ is nearly constant over the subbands of each of our images, and thus need not be sent for each subband (although an advantage of the current implementation is that it can handle contamination by additive noise that is not spectrally white). Similarly, the α parameter could probably be estimated rather than sent for each subband.

We are working to incorporate the probability model into other wavelet-based applications, such as image enhancement and texture synthesis. Given the role of noise in the conditional distribution (see figure 5), the model should prove useful in blind noise removal. This type of application, however, is likely to require an aliasing-free representation, such as an overcomplete [e.g., 25] or

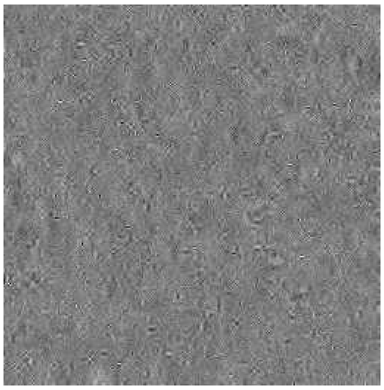

Figure 14: A hard-to-code image for EPWIC-2, synthesized by drawing coefficients from the marginal densities of each subband. The density parameters are those that best fit the subbands of the Lena image.

adaptive [e.g., 10, 4] basis.

7 Acknowledgments

We would like to thank Dr. Ruzena Bajcsy and the members of the GRASP lab for their support during the course of this work.

References

[1] E H Adelson, E P Simoncelli, and R Hingorani. Orthogonal pyramid transforms for image coding. In *Proc. SPIE*, volume 845, pages 50–58, Cambridge, MA, October 1987.

[2] R W Buccigrossi and E P Simoncelli. Progressive wavelet image coding based on a conditional probability model. In *ICASSP*, Munich, Germany, April 1997. To Appear.

[3] C Chrysafis and A Ortega. Efficient context-based entropy coding for lossy wavelet image coding. In *Data Compression Conference*, Snowbird, Utah, March 1997.

[4] R Coifman and W Wickerhauser. Best-adapted wave packet bases. Technical report, Numerical Algorithms Research Group, Dept of Math, Yale University, 1990.

[5] D L Donoho. Nonlinear wavelet methods for recovery of signals, densities, and spectra from indirect and noisy data. In I. Daubechies, editor, *Proc Symp Appl Math*, volume 47, pages 173–205, Providence, RI, 1993.

[6] D J Field. What is the goal of sensory coding? *Neural Computation*, 6:559–601, 1994.

[7] H Gharavi and A Tabatabai. Sub-band coding of digital images using two-dimensional quadrature mirror filtering. In *Proc. of SPIE*, volume 707, pages 51–61, 1986.

[8] D. Heeger and J. Bergen. Pyramid-based texture analysis/synthesis. In *Proc. ACM SIGGRAPH*, August 1995.

[9] F. Kossentini, W. Chung, and M. Smith. A jointly optimized subband coder. *IEEE Trans Image Processing*, 5(9):1311–1323, September 1996.

[10] Stephane Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signal Processing*, December 1993.

[11] Stephane G. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. PAMI*, 11:674–693, July 1989.

[12] B A Olshausen and D J Field. Natural image statistics and efficient coding. *Network: Computation in Neural Systems*, 7:333–339, 1996.

[13] A Pentland and B Horowitz. A practical approach to fractal-based image compression. In A B Watson, editor, *Digital Images and Human Vision*. MIT Press, 1993.

[14] A P Pentland, E P Simoncelli, and T Stephenson. Fractal-based image compression and interpolation, 1992. U.S. Patent Number 5,148,497, filed 2/14/90, issued 9/15/92.

[15] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery. *Numerical Recipes in C - Second Edition*. Cambridge University Press, Cambridge, MA, 1992.

[16] R Rinaldo and G Calvagno. Image coding by block prediction of multiresolution subimages. *IEEE Trans Image Processing*, July 1995.

[17] D L Ruderman and W Bialek. Statistics of natural images: Scaling in the woods. *Phys Rev. Letters*, 73(6), August 1994.

[18] A Said and W A Pearlman. An image multiresolution representation for lossless and lossy compression. *IEEE Trans. Image Proc*, 5(9), September 1996.

[19] E Schwartz, A Zandi, and M Boliek. Implementation of compression with reversible embedded wavelets. In *Proc SPIE*, 1995.

[20] Jerome Shapiro. Embedded image coding using zerotrees of wavelet coefficients. *IEEE Trans Signal Processing*, 41(12):3445–3462, December 1993.

[21] E P Simoncelli. Orthogonal sub-band image transforms. Master's thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, Cambridge, MA, May 1988. Also available as MIT Media Laboratory Vision and Modeling Technical Report #100.

[22] E P Simoncelli and E H Adelson. Subband transforms. In John W Woods, editor, *Subband Image Coding*, chapter 4, pages 143–192. Kluwer Academic Publishers, Norwell, MA, 1990.

[23] E P Simoncelli and E H Adelson. Noise removal via bayesian wavelet coring. In *Third Int'l Conf on Image Processing*, pages 379–383, Lausanne, Switzerland, September 1996.

[24] E P Simoncelli and R W Buccigrossi. Progressive wavelet image compression using linear inter-band magnitude prediction. In *Fourth Int'l Conf on Image Processing*, Santa Barbara, October 1997. Submitted 2/1/97.

[25] E P Simoncelli, W T Freeman, E H Adelson, and D J Heeger. Shiftable multi-scale transforms. *IEEE Trans. Information Theory*, 38(2):587–607, March 1992. Special Issue on Wavelets.

[26] Martin Vetterli. Multi-dimensional sub-band coding: Some theory and algorithms. *Signal Processing*, 6(2):97–112, February 1984.

[27] J W Woods and S D O'Neil. Subband coding of images. *IEEE Trans. Acoust. Speech Signal Proc.*, ASSP-34(5):1278–1288, October 1986.

[28] G Wornell. Wavelet-based representations for the $1/f$ family of fractal processes. *Proc. IEEE*, September 1993.

[29] X Wu and J Chen. Context modeling and entropy coding of wavelet coefficients for image compression. In *ICASSP*, Munich, April 1997.

[30] S Zhu, Y Wu, and D Mumford. Filters, random fields and maximum entropy (FRAME) – towards the unified theory for texture modeling. In *IEEE Conf. Computer Vision and Patt Rec*, June 1996.